УДК 519.8

М. Милош – кандидат технических наук, заведующий кафедрой программного обеспечения и баз данных

Люблинский политехнический университет (Польша)

ТЕХНОЛОГИЯ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ В СОВРЕМЕННЫХ МЕТОДАХ АНАЛИЗА ДАННЫХ

АННОТАЦИЯ

На сегодняшний день ученые и предприятия загружены большим количеством данных. Данные сами по себе бесполезны без современных методов извлечения из них знаний. Эти методы названы оперативным анализом данных. Статья посвящена технологии оперативного извлечения информации при анализе данных и получения из них знаний.

M. Milosz – candidate of technical sciences, head of the Software Engineering and Database Systems department

Lublin University of Technology (Poland)

DATA MINING AS A MODERN METHOD OF DATA ANALYSIS

ABSTRACT

Nowadays scientists and companies are over flooded with a huge amount of data. Those data are useless without proper techniques, extracting the required knowledge from them. This kind of techniques is called as data mining. This paper describes the usage of data mining for data analysis and knowledge obtaining.

1. Data, information and knowledge

Information technology has made possible to collect huge amount of different date in databases. Databases usually are sources of information. Using the simple transformation (calling database query) it is possible to convert data into information. More complicate problem is knowledge acquisition from data.

Knowledge indeed is hard to formalise, serve and to manage. The reason for it is that knowledge is very unstable. It is not easy to update, gain and measure it. That is why knowledge management (for example with computer system assistance) is harder than information or data assistance. It is shown on knowledge triangle (Fig. 1).

Fig. 1 illustrates knowledge, which is superior unit against data and information. Simultaneously data and information are bases of knowledge. Data are a set of severe, disordered facts, which are not connected with themselves. Classified and categorized facts are called information. These are areas of management computer systems (MCS) operations. Knowledge management computer systems (KMCS), however, are one jump ahead of MCS. They require more complex tools, which can partly replace and assist human. They gain, accumulate and pass on knowledge. Knowledge is processed, selected, related information. This kind of systems are separate group because of a way of using knowledge, used tools and technologies and character of knowledge indeed. Ability to use knowledge in practice is known as wisdom. However wisdom can be processed only by human. Inappropriate or incompetent usage of knowledge leads to loss of usefulness. As a result it changes knowledge in information or data again.

According to Nonaka and Takeuchi [17] model knowledge can be divided into two types: *explicit* and *tacit* knowledge. Explicit knowledge is freely available.

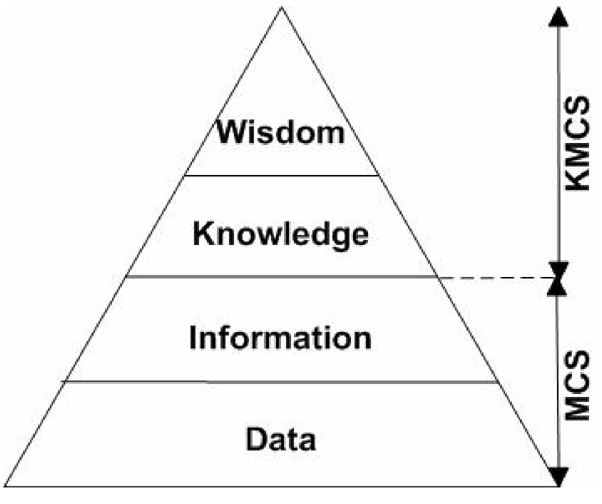

Fig. 1. Knowledge, data, information and wisdom and computer systems.
Personal elaboration on the basis [4]

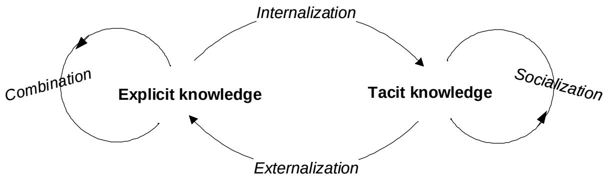

Fig. 2. Process of knowledge conversion.
Personal elaboration on the basis [4]

It can be reached in books, reports and other ways of notation [8]. Nevertheless tacit knowledge is hard to deal with. It is subject of researches and worries IT specialists. This kind of knowledge can be available in employee’s minds, skills, experiences, intuitions. Making it freely available depends on goodwill of owner. On the other hand this kind of knowledge is the most precious because it is practical knowledge. It can be passed over through descriptions, stories, informal notes.

Knowledge management should concern both of knowledge types. As a result it will be easer to gain knowledge from such sources as databases or sharing it by workers. Besides interesting information and knowledge will be found in simpler way. There are four transformations between knowledge types (fig. 2).

Socialization [17] is a transformation of tacit knowledge into tacit one. It refers to passing on modified, competed knowledge. Socialization has an effect on workers minds. It can be found in conversations, discussions, using forums and experts advising.

Externalization [17] is the process, where tacit knowledge is transformed into explicit one. It occurs during formation of explicit knowledge. Process of externalization consists of recording knowledge using electronic documents, databases. computer systems and knowledge repositories.

Transforming explicit knowledge into explicit one is known as ***combination*** [17]. It is processing of earlier recorded information, cataloguing or rewriting them.

Internalization [17] is a transformation of explicit knowledge into tacit one. The best example is learning process.

Task of KMCS is to create new knowledge in process of externalization and to preserve it and adjust to clients needs in process of combination. Besides system should make possible to gain knowledge (internalization) and to exchange it (socialization).

2. Knowledge management computer system

KMCS consists of (Fig. 3):

- The part responsible for knowledge extraction – basic element of the system. It is responsible for gaining, modelling and interpretation of knowledge. Knowledge extraction makes possible functioning of whole system. Area of knowledge extraction is discussed in next parts of this paper.

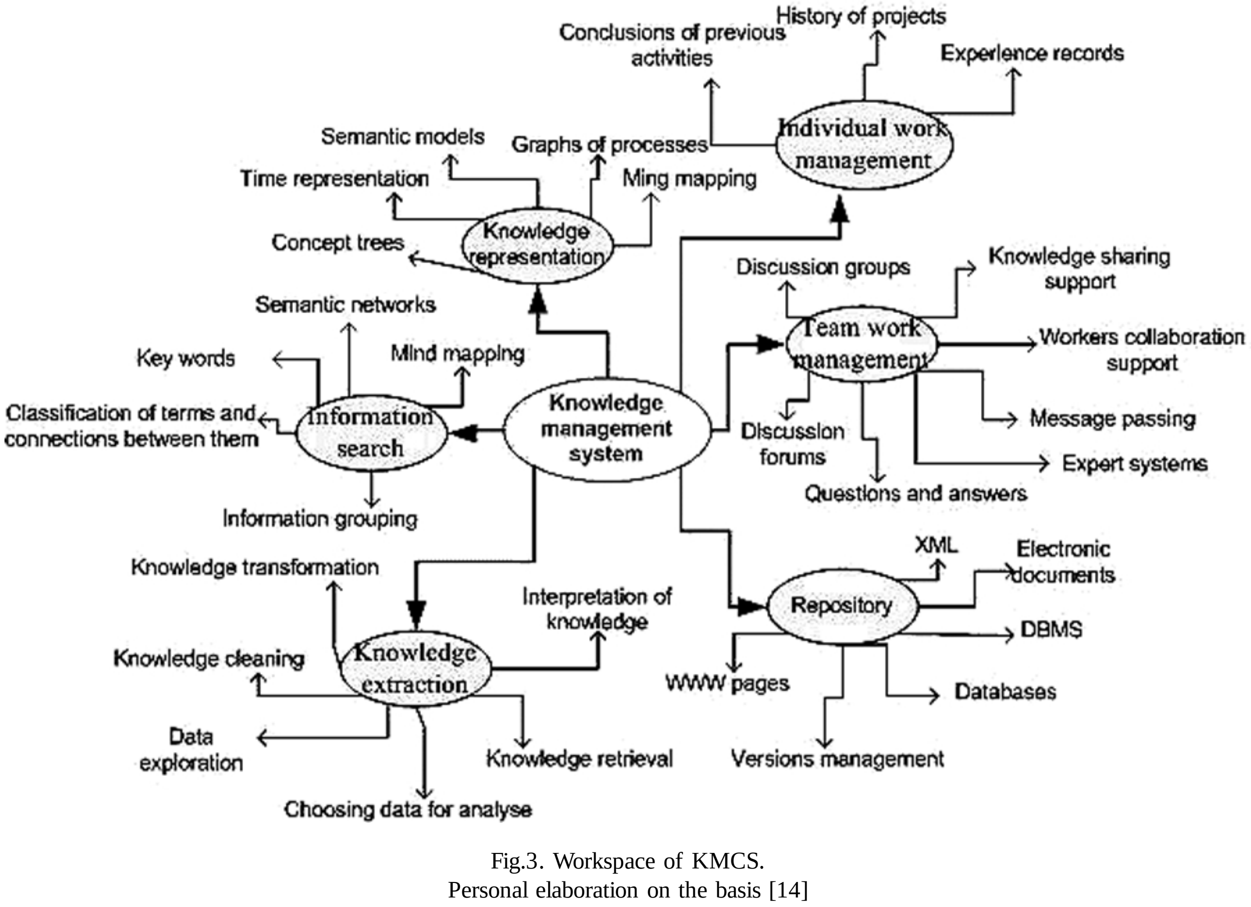

Fig.3. Workspace of KMCS.
Personal elaboration on the basis [14]

- The repository [13] – an area to keep knowledge and information, which that knowledge is made of. In repository there are implemented such mechanisms as files management and their XML representation as well as databases management systems.
- Information search module. There can be implemented different mechanisms, depending on needs. The simplest and simultaneously little effective solutions use key-words. More sophisticated methods make use of mind mapping or semantic networks. From system functionality point of view information searching is important, but difficult issue. Computer does not posses an intuition assisting it in distinguishing meaning of words. To make this process automatic there was found concepts classification and connections between concepts classifications. As a result it is possible to automatically group information and analyse documents.
- Individual work management [6]. It is consists of recorded historical projects and any other experiences, which might be useful in future. Analyzing interesting events from past leads to creating knowledge. This knowledge helps to draw a conclusion and gives a good chance of success.
- Team work management [6] is a group of all solutions motivating and assisting in sharing knowledge and workers collaboration. Some of them are: expert systems, discussion forums, documents approving, questions and answers systems.
- The representation [13] – an area responsible for knowledge presenting in the system. There can be found such mechanisms as mind mapping, semantic networks and models or concept trees. They are used for an enterprise modelling. Right model should consider important issues but pass over unnecessary details.

KMCS use generally known mechanisms as business intelligence, expert systems, documents management, workers collaboration support systems, decision support systems, knowledge bases, alerts systems, data marts, project management, reports generating, customer relationship management [1]. The most important elements however are gaining, extracting, transforming and passing on the knowledge [15]. It is important because of knowledge live circle (fig. 4). Knowledge live circle must be present in every knowledge management system. It is divided into three parts: creating, validation and integration of knowledge. There is no benefits without property knowledge manipulation. This is because of knowledge character.

There can be found following actions in most knowledge circulation schemas [1]:

- knowledge acquisition from exterior sources,
- new knowledge creating and modifying existing one,
- choosing interesting knowledge and transforming it to make it useful,
- modelling and recording knowledge,
- ability to make use of chosen knowledge.

3. Automatic knowledge acquisition – data mining

There is an important question concerning knowledge management. This question is “how to get this knowledge?”. Science of today are rained with various information. Databases grow fast. However if having this huge piece of information means taking advantage of them? Surely. There is need to use additional tools to make even the best database useful for knowledge management. Knowledge acquisition can be execute through proper cooperation culture [6] and implementation of solution, which can improve externalisation and combination processes.

Technology of knowledge extraction in knowledge base is called data mining [10]. Data mining makes automatic knowledge extraction possible. It is used to find interesting connections, relations, patterns, answers to decision questions. Direct target of data mining is often prediction. Predictive data mining can be used to market situation, clients targets or decisions results. In every case it can be helpful to keep market position and to increase development chance.

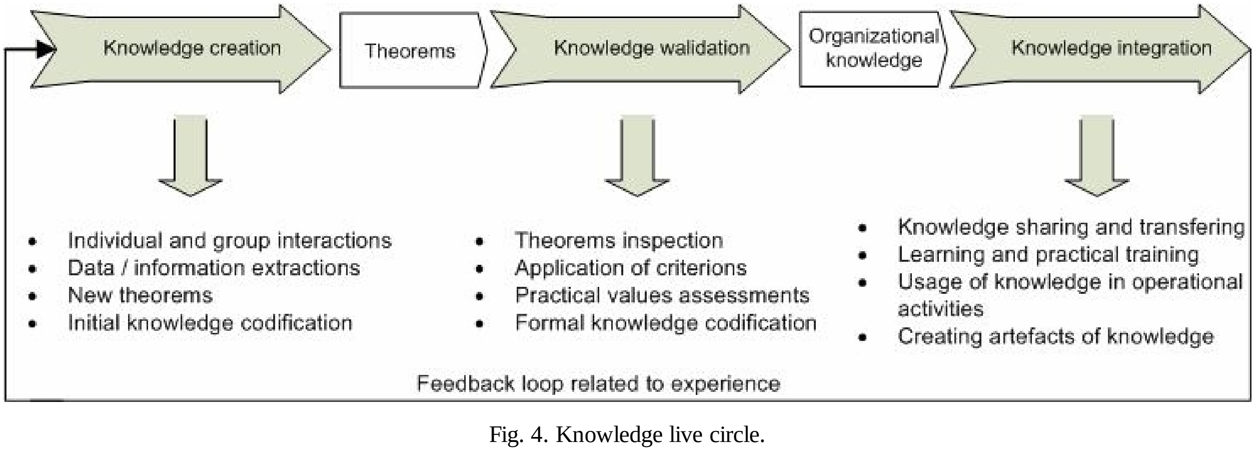
Knowledge creation
Theorems
Knowledge walidation
Organizational knowledge
Knowledge integration
Individual and group interactions
Data / information extractions
New theorems
Initial knowledge codification
Theorems inspection
Application of criterions
Practical values assessments
Formal knowledge codification
Knowledge sharing and transfering
Learning and practical training
Usage of knowledge in operational activities
Creating artefacts of knowledge
Feedback loop related to experience

Fig. 4. Knowledge live circle.
Personal elaboration on the basis [13]

A good example of data mining is seeking huge databases in supermarkets [18]. Those databases arise from bar codes. Bar codes were found for supplying purpose, to determine demand for articles. However in that database can be found knowledge concerning clients and their needs. Based on that knowledge, managers can adjust market strategies to clients needs. That knowledge might concern such issues as types of products bought together, products which can be seasonal, contents of small shopping baskets. Interesting relations are relations, which are not obvious. They are rather surprising.

Another example is cell telephoning. Personal clients data can give a lot of interesting information as their likes, demands or feelings concerning specific products. The assertion that, for example, the best clients are middle-age men or that young women are people who the most often change phone. Having that knowledge it is easer to make a marketing decisions. Above-mentioned examples includes elements of decision making systems and customer relationship management. They are in KMCS activities area, just like efficiency and satisfaction analyse. Using current and archive employee data there can say a lot about personal policy of corporation. For example if many workers have left office taking their knowledge for last few years, personal policy need to be thought over.

Data mining technology can make company management easier. It can be done through analysing and conducting from information resources. Data mining makes possible to check any solutions, helps to introduce new ideas, fasts reactions on market changes [10]. It also predicts clients loyalty, helps finding their needs, supplies information about them. Another use is to make data coherent and uniform. It may be helpful in defining and understanding corporation processes, making reports and making knowledge available in one source.

Knowledge extraction was found because of much interest in information recovery. Data marts were found to describe activity areas in the course of time [18]. But there are disadvantages of this technology. One of them is a necessity to constant control and maintenance of an analyst. An analyst needs to define hypothesises and to serve a query. It needs to be remember about demand for computational power. Databases often include millions of records, so processing them can be time-consuming [9].

As it turned out, data mining is a mile stone of knowledge discovery. Knowledge discovery is automatic. It needs only small assistance of an analyst, who is responsible for pointing at searching area. There are used models built of known, archive data. Those models, filled up with current data, makes prediction possible. This kind of models predict for ex. market behaves. Writing new data into old model may be successful. Experiences form the past can be helpful in answering questions concerning the future. This is known as prediction. There is possibility to predict same interesting variables using actual, current values of that variables. Use of historical data makes predicting possible and efficient. For example sales balance sheets for last few years can be helpful in autumn sales of wines prediction or in finding the most popular sweets in next spring.

Answers on those questions can be used only one and, unnecessary, deleted. However sometimes it is profitable to keep them. Answers on important questions, put in knowledge base, can make easier to solve another decision problems and to analyse new ideas.

4. Data mining techniques

There are many types of data miming techniques. This diversity is caused of demand for different types of actions, different data imperfection or hardware abilities.

The most known and popular techniques use statistical methods, like averages, charts, histograms, percentage rate. Additionally, to make results more possible, there is used of data distinction between continuous and non-continuous variables. Averaging [11] is used for continuous variables, and voting [11] – otherwise. Another technique are regression methods – methods using mathematical functions to modelling and predicting different processes and approximating results.

Often used techniques are based at models, which use archive data. Those models can be used to construction new ones. They also can be a form to filled with new data. A technique called bagging [11] is a good example. Predicting is based on few models the same type, but filled with different data. In special cases used models might be different types, but filled with the same data. Another solution is boosting [11]. It consist in construction collective model using different models built with interesting data. Additionally weights can be used to make predicting more accurate.

Another group of algorithms and techniques are aggregation methods. They introduce automation into data classification processes. This group consists of, among other things, clustering, classification and association [9]. Clustering is grouping records by some criterions. Those criterions are automatically generated. Clustering makes possible to find hidden irregularity in database, because it discloses records, which match no group. Classification is grouping records by given criterions, specified by user or analyst. Decision trees and semantic networks are mechanisms, which can be used in classification process. Association is finding elements concerned or similar to other elements or events. Often used rule is “if element A is an component of some event, element B is also component of the same event in X% of cases”.

Nearest neighbour algorithm [9] is suitable when data are incomplete. Lacking values of feature can be estimated on the basis of value of this feature for different elements. Those elements should be at the same range or conditions as the searching one. Elements can be numbered among one group by some common features. The best “neighbours” are elements, which show close likeness by specified class of patterns.

Neural networks are ones of the most advised techniques of data mining. They are complex techniques created on the model of neurological functions of a brain. Constructing model is describing through many coefficients. That is why it is so complicated.

Commonly used techniques are visualisation techniques. They are popular because of easily of usage. They use graphical methods improving trends identification and behaves hidden in databases. The most known example is distinguishing technique [11]. It concerns choose hole data sets or individual points on a chart and identification their common features, characteristics, dependencies.

Pattern detection technique written as a text (documents) is called text mining [11]. This kind of data are difficult to analyse, because it is hard to standardise them, and heir structure is not specified. Documents analyse is based on searching given words or expressions. Text mining makes easier to classify documents and extract knowledge. They also determine importance of documents and place them into repositories.

5. Data mining as way of knowledge extraction

According specialists [12], terms data mining and knowledge extraction are unequal. Knowledge extraction is a term more general. It applies to hole knowledge retrieval process. Data mining, known as data exploration, applies only to techniques and tools. However, especially in common language, both of those terms are used interchangeably. This interchangeability, for this paper needs, is acceptable.

Knowledge extraction process is proceeded as follows (fig. 5):

1. Data choice (selection) to separate an area for analysis. That chosen data must meet given demands, conditions, for example time intervals or place of client residence.

2. Data transformation to a form, which make analyse possible.

3. Data exploration (seemly data mining) – choice and usage of algorithms, techniques and software adapted to extract information, patterns and knowledge. If decided choice turn out to be wrong, there is possibility to back to data choice (point nr. 1).

4. Presenting and interpretation of results of searching the most interesting knowledge, as well as assessment, cleaning and filtering of knowledge.

Knowledge extraction process entirely can represent process of transformation data info knowledge and even wisdom. Data, coming from different sources, first lend in database. Those data can concern, for example, personal details of clients, shopping (bar codes), likes and interests of clients. Accumulated this way data usually are large sized. Simultaneously such cluster of not-related data is useless, and manual segregation is impossible. However according to fig. 6 this kind of data can be transformed in proper way. Suitable columns choice, conditions applying and range determining make possible to transform data into information. Classified and categorized data become information (fig. 5). Those information are not directly useful yet. However, they became systematised knowledge after interpretation, connecting with specific problems, processes or persons. That knowledge can be used by human to create correct project, to find better solutions, to make right decisions. Usage of knowledge in practise is a wisdom which, however, can only be share of human. Thus, data mining, as knowledge discovering process, is technology of extraction and creation new knowledge. As a result it represents important part of knowledge management computer systems, because it is responsible for acquisition and accumulation of knowledge, especially from databases.

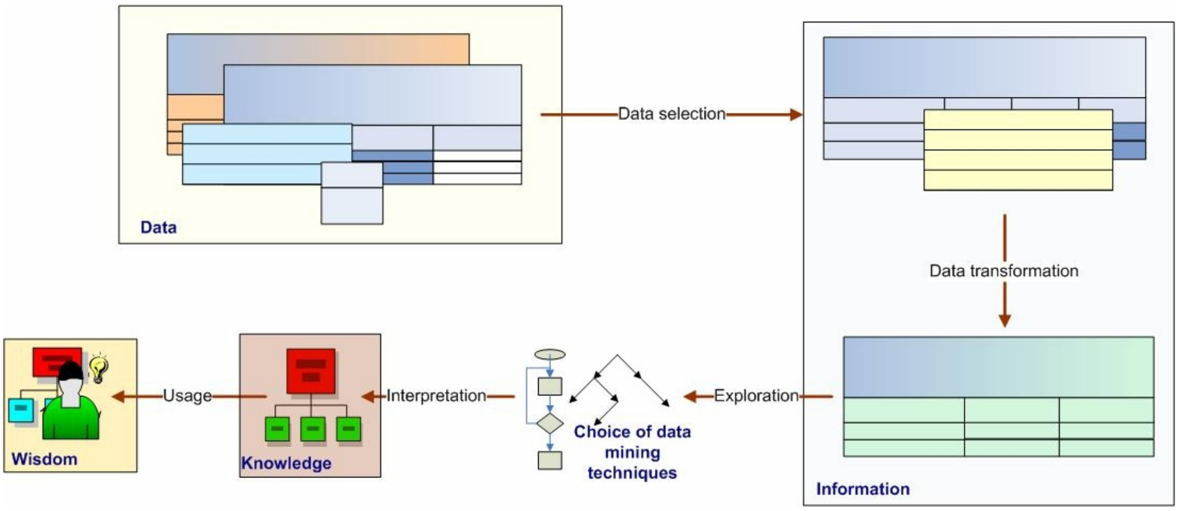

Fig. 5. Knowledge extraction process.
Personal elaboration on the basis [5]

7. Conclusions

Data mining entirely cannot be regarded as technology sufficient to make KMCS. However it is one of its basic elements: choosing and integrating knowledge. It makes possible to order, to record and to understand excessive amount of information. It is impossible, however, using it to encourage employee to share with knowledge and to support them to collaboration between themselves..

References

1. Plechawska M., Miłosz M.: Data Mining In Knowledge Management Systems. In: Studies&Proceedings of Polish Association for Knowledge Management, vol. 9, Bydgoszcz, 2007, pp. 87-97.
2. Miłosz M., Plechawska M.: System zarządzania wiedza w elektrociepłowni Megatem-EC Lublin – koncepcja wstępna. red. Miłosz M., Muryjas P.: Varia informatica - teoria i praktyka 2005, PTI, Lublin 2005.
3. Plechawska M., Wójcik M.: Problemy i ryzyko przy wdrazaniu systemow zarzadzania wiedzą (na przykładzie przedsiębiorstwa – elektrociepłowni Megatem-EC). Ryzyko przedsięwzięć informatycznych, Politechnika Szczecińska, Szczecin 2005.
4. Grudzewski W., Hejduk I.: Zarządzanie wiedzą w przedsiębiorstwach. Wydawnictwo „Difin", Warszawa 2004.
5. Marcin Domański: Data mining, obecny stan i możliwości rozwoju. In żynieria gier komputerowych. Materiały konferencyjne III Ogólnopolskiej Konferencji Gier Komputerowych. Wydawnictwo Akademii Podlaskiej, Siedlce 2006.
6. Leszek Panasiewicz: Systemy zarządzania wiedzą – przegląd metodologiczny. Systemy informatyczne zarządzania – od teorii do praktyki. PWN, Warszawa 2006.
7. Robert Rusielik: Atrakcyjność inwestowania w metody zarządzania wiedza w agrobiznesie. Studia i materiały polskiego stowarzyszenia zarządzania wiedza.
8. Zbigniew J. Klonowski: Systemy informatyczne gospodarowania wiedza. Studia i materiały polskiego stowarzyszenia zarządzania wiedza.
9. Michał Gulczynski: Techniki „odkrywania wiedzy" (data mining) oraz ich zastosowania. Studia i materiały polskiego stowarzyszenia zarządzania wiedza.
10. Kurt Thearling: Data Mining and Analytic Technologies, http://www.thearling.com/.
11. Portal StatSoft, Techniki zgłębiania danych (data mining) http://www.statsoft.pl/textbook/stathome_stat.html?http%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstdatmin.html.
12. Muraszkiewicz M., Eksploracja danych dla telekomunikacji http://www.icie.com.pl/ARCHIVE/dm_tel.doc.
13. Staniszkis E., Nowicki B.: Zarządzania wiedzą w procesie przygotowywania propozycji projektów funduszy strukturalnych w oparciu o platformę ICONS. http://www.rodan.pl/pdf/SF_pl_0215.pdf.
14. Krzysztof Regulski: Komputerowe systemy zarządzania wiedzą i ewidencji kapitału intelektualnego, http://www.regulski.padlock.pl/pliki/seminarium.pdf.
15. Hemamalini Suresh BE , MBA: Knowledge Management - The Road Ahead for Success http://knowledgemanagement.ittoolbox.com/pub/HS090302.pdf.
16. Tiffany Blaine: Knowledge Management Overview http://knowledgemanagement.ittoolbox.com/pub/km_overview.htm.
17. Richard G. Best, Sylvia J. Hysong, Charles McGhee, Frank I. Moore, Jacqueline A.: Pugh An Empirical Test of Nonaka's Theory of Organizational Knowledge Creation http://www.weleadinlearning.org/rboct03.htm.
18. Chris T.: Data Warehouse: Supporting Customer Relationship Management, First Edition, published by Pearson Education, Inc., Pretentice Hall.